INSTITUTE OF INFORMATICS
SLOVAK ACADEMY OF SCIENCES

A State-of-the-Art Local Training Methods in Federated Learning

Michal Staňo, Ladislav Hluchý

INSTITUTE OF INFORMATICS
SLOVAK ACADEMY OF SCIENCES, BRATISLAVA SLOVAKIA

IEEE 23rd International Symposium on Computational Intelligence and Informatics (CINTI 2023)
November 20, 2023

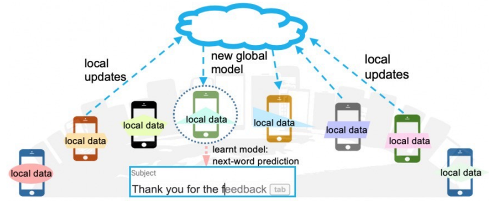
Outline of the Talk

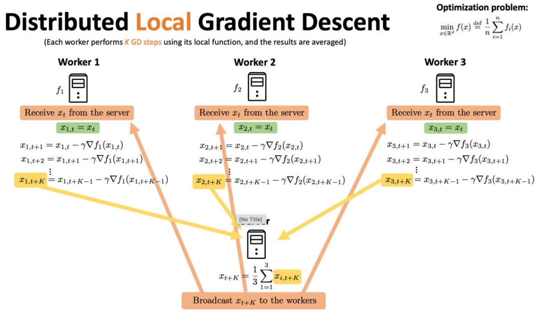
1. What is Federated Learning?

2. What is Local Training

3. Brief History of Local Training

4. What does Local Training do?





Distributed Local Gradient Descent

(Each worker performs K GD steps using its local function, and the results are averaged)

Optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Receive x_t from the server

$x_{1,t} = x_t$

$x_{1,t+1} = x_{1,t} - \gamma\nabla f_1(x_{1,t})$

$x_{1,t+2} = x_{1,t+1} - \gamma\nabla f_1(x_{1,t+1})$

$\vdots$

$x_{1,t+K} = x_{1,t+K-1} - \gamma\nabla f_1(x_{1,t+K-1})$

Receive x_t from the server

$x_{2,t} = x_t$

$x_{2,t+1} = x_{2,t} - \gamma\nabla f_2(x_{2,t})$

$x_{2,t+2} = x_{2,t+1} - \gamma\nabla f_2(x_{2,t+1})$

$\vdots$

$x_{2,t+K} = x_{2,t+K-1} - \gamma\nabla f_2(x_{2,t+K-1})$

Receive x_t from the server

$x_{3,t} = x_t$

$x_{3,t+1} = x_{3,t} - \gamma\nabla f_3(x_{3,t})$

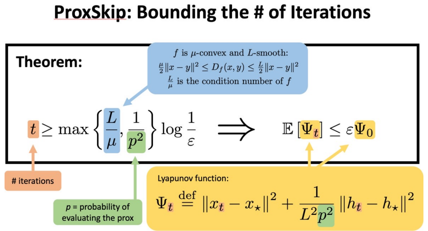
$x_{3,t+2} = x_{3,t+1} - \gamma\nabla f_3(x_{3,t+1})$

$\vdots$

$x_{3,t+K} = x_{3,t+K-1} - \gamma\nabla f_3(x_{3,t+K-1})$

$$x_{t+K} = \frac{1}{3}\sum_{1=1}^{3} x_{i,t+K}$$

Broadcast x_{t+K} to the workers



ProxSkip: Bounding the # of Iterations

Theorem:

$$t \geq \max\left\{\frac{L}{\mu}, \frac{1}{p^2}\right\}\log\frac{1}{\varepsilon} \quad\Longrightarrow\quad \mathbb{E}\left[\Psi_t\right] \leq \varepsilon\Psi_0$$

f is μ-convex and L-smooth:

$\frac{\mu}{2}\|x-y\|^2 \leq D_f(x,y) \leq \frac{L}{2}\|x-y\|^2$

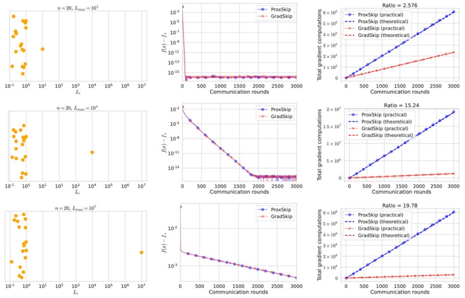
$\frac{L}{\mu}$ is the condition number of f

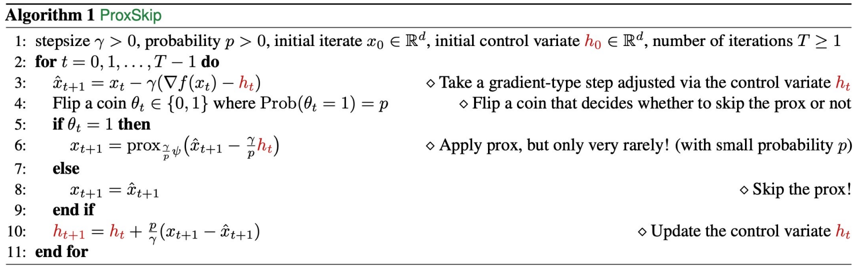
iterations

p = probability of evaluating the prox

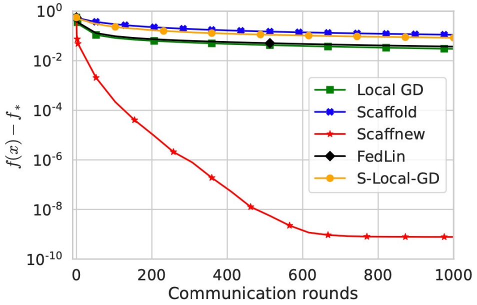
Lyapunov function:

$$\Psi_t \overset{\text{def}}{=} \|x_t - x_\star\|^2 + \frac{1}{L^2p^2}\|h_t - h_\star\|^2$$





```
Algorithm 1 ProxSkip
stepsize $\gamma > 0$, probability $p > 0$, initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$
for $t = 0, 1, \ldots, T-1$ do
    $\hat{x}_{t+1} = x_t - \gamma(\nabla f(x_t) - h_t)$        ◇ Take a gradient-type step adjusted via the control variate $h_t$
    Flip a coin $\theta_t \in \{0,1\}$ where $\text{Prob}(\theta_t = 1) = p$        ◇ Flip a coin that decides whether to skip the prox or not
    if $\theta_t = 1$ then
        $x_{t+1} = \text{prox}_{\frac{\gamma}{p}\psi}(\hat{x}_{t+1} - \frac{\gamma}{p}h_t)$        ◇ Apply prox, but only very rarely! (with small probability $p$)
    else
        $x_{t+1} = \hat{x}_{t+1}$        ◇ Skip the prox!
    end if
    $h_{t+1} = h_t + \frac{p}{\gamma}(x_{t+1} - \hat{x}_{t+1})$        ◇ Update the control variate $h_t$
end for
```

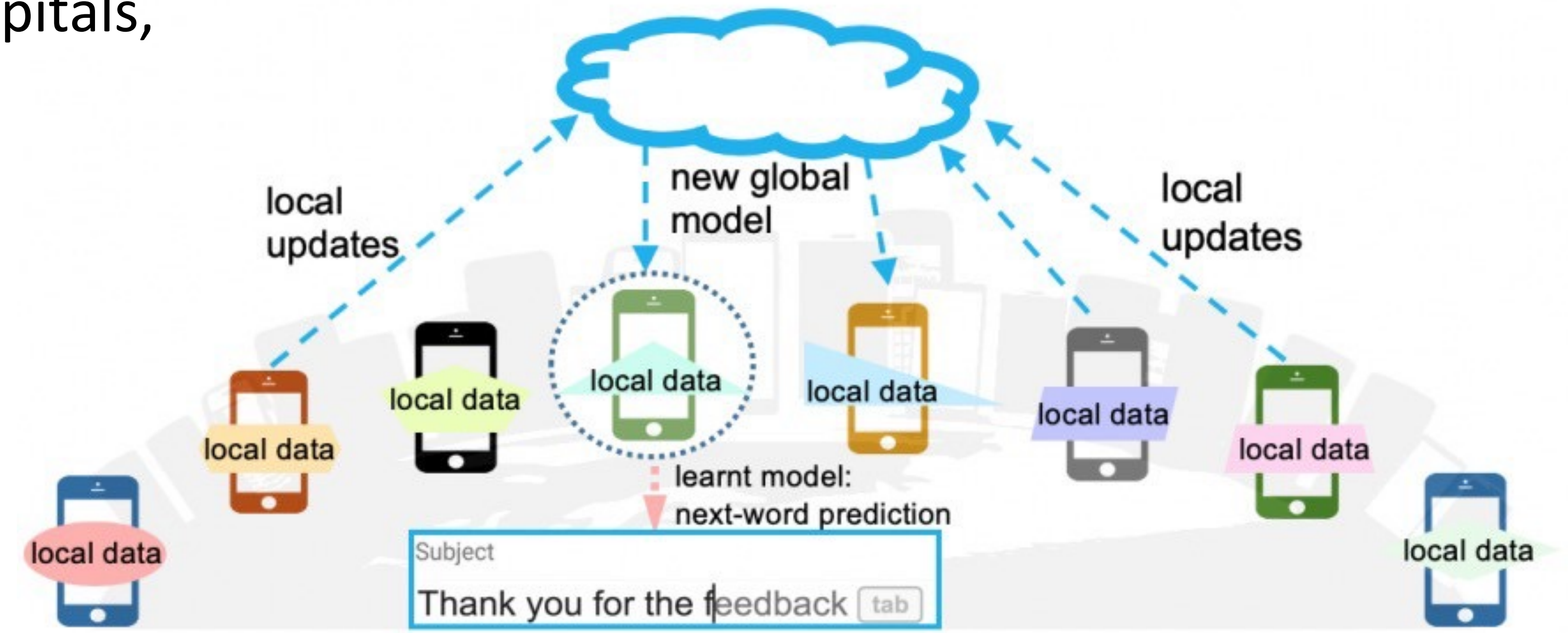


(c) theoretical hyper-parameters

Part 1
What is Federated Learning?

The First Federated Learning App: Next-Word Prediction?

Federated Learning is a collaborative machine learning from private data stored across a (large) number of clients/devices (e.g., hospitals, phones, banks)



Part 2
What is Local Training?

Local Training

A. Gradient Descent

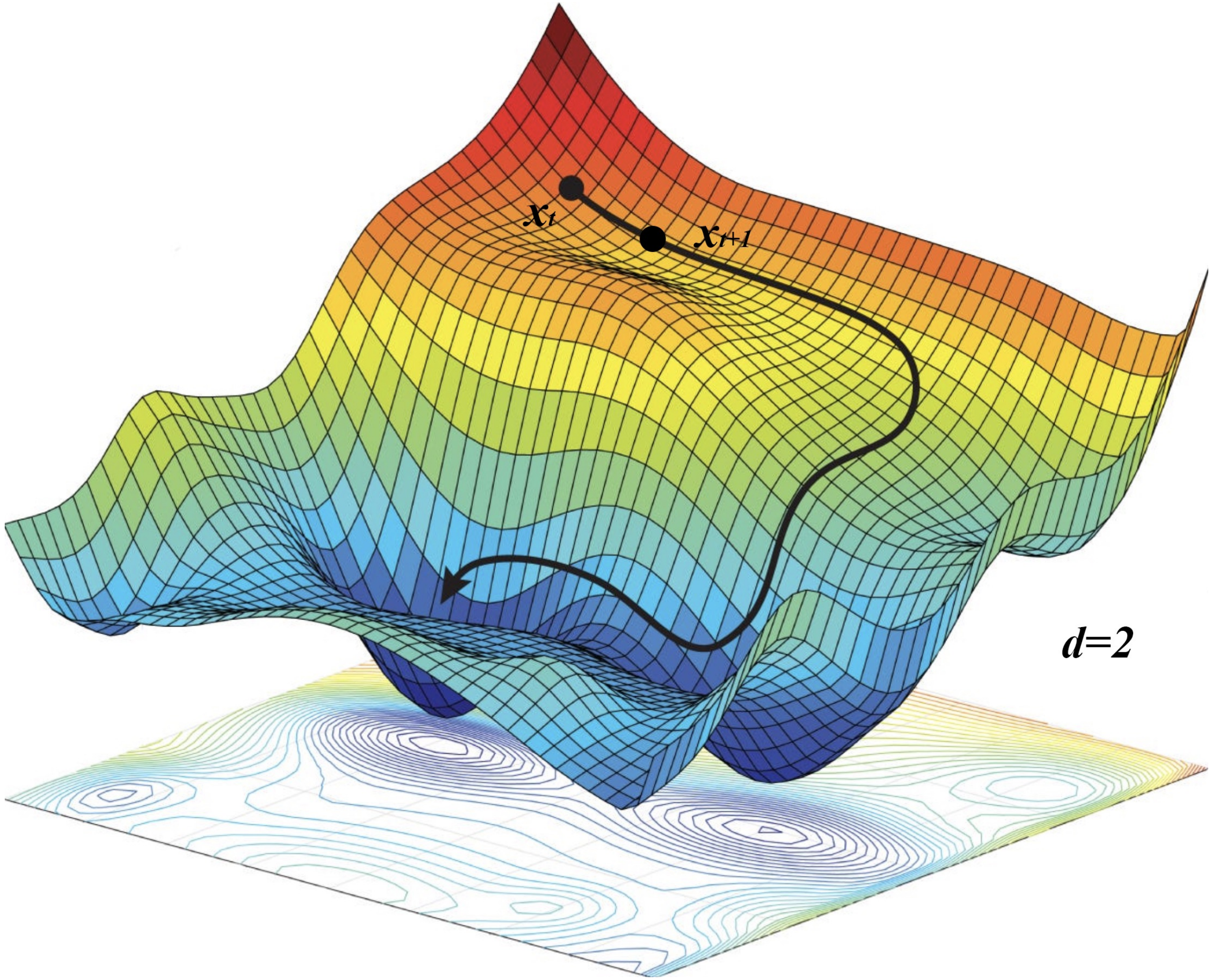
B. Distributed Gradient Descent

C. Distributed Local Gradient Descent

Gradient Descent

$$\min_{x \in R^d} f(x)$$

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$



Distributed Gradient Descent

$$\min_{x \in R^d} f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Distributed Gradient Descent

$$\min_{x\in\mathbb{R}^{d}} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

# devices / machines

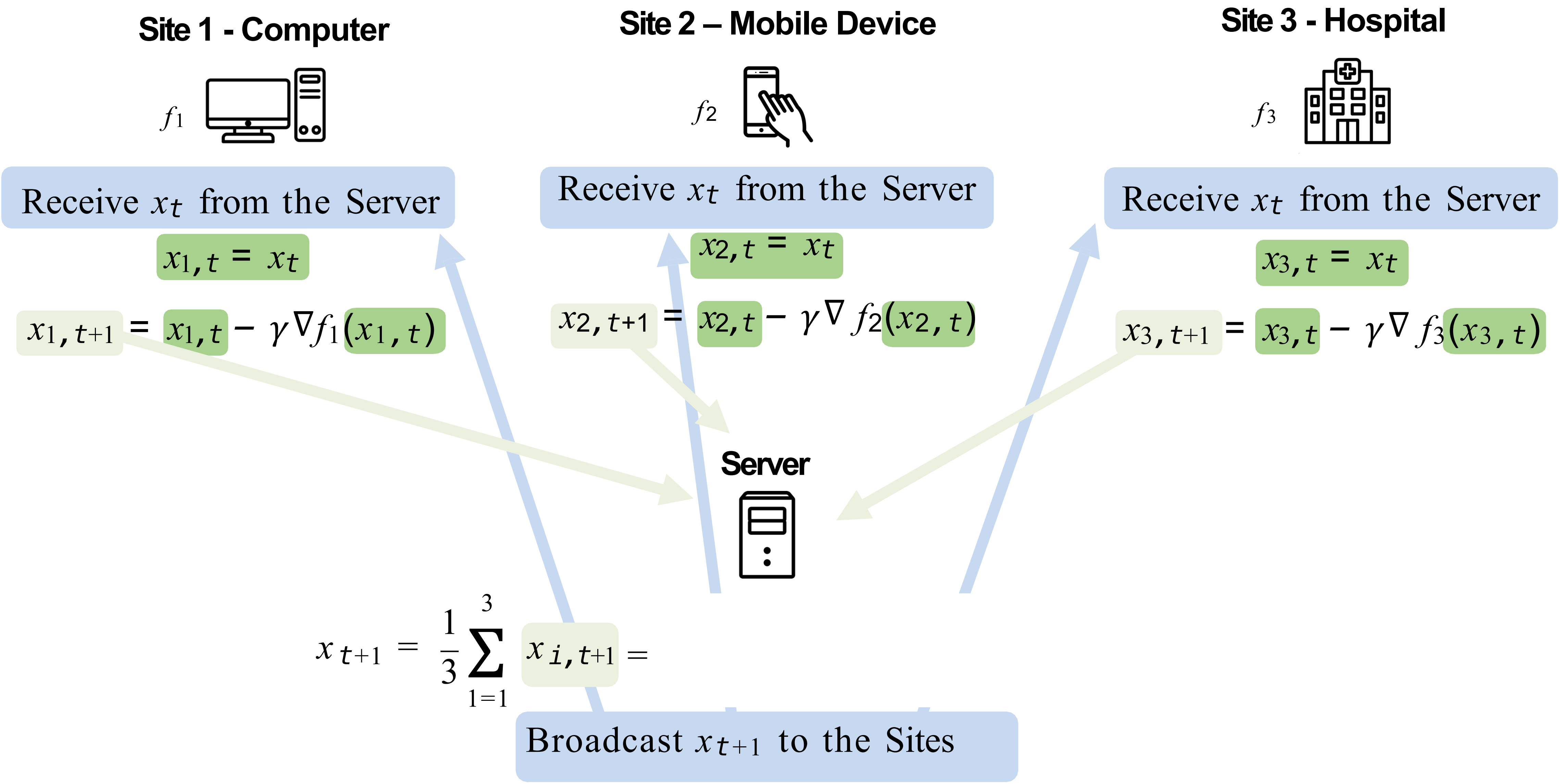
# model parameters / features

Loss on local data D_i stored on device i

$$f_i(x) = \mathbb{E}_{\xi\sim D_i} f_{i,\xi}(x)$$

The datasets $D_1,\ldots,D_n$ can be arbitrarily heterogeneous

Distributed Gradient Descent



Site 1 - Computer

f_1

Receive x_t from the Server

$x_{1,t} = x_t$

$x_{1,t+1} = x_{1,t} - \gamma \nabla f_1(x_{1,t})$

Site 2 – Mobile Device

f_2

Receive x_t from the Server

$x_{2,t} = x_t$

$x_{2,t+1} = x_{2,t} - \gamma \nabla f_2(x_{2,t})$

Site 3 - Hospital

f_3

Receive x_t from the Server

$x_{3,t} = x_t$

$x_{3,t+1} = x_{3,t} - \gamma \nabla f_3(x_{3,t})$

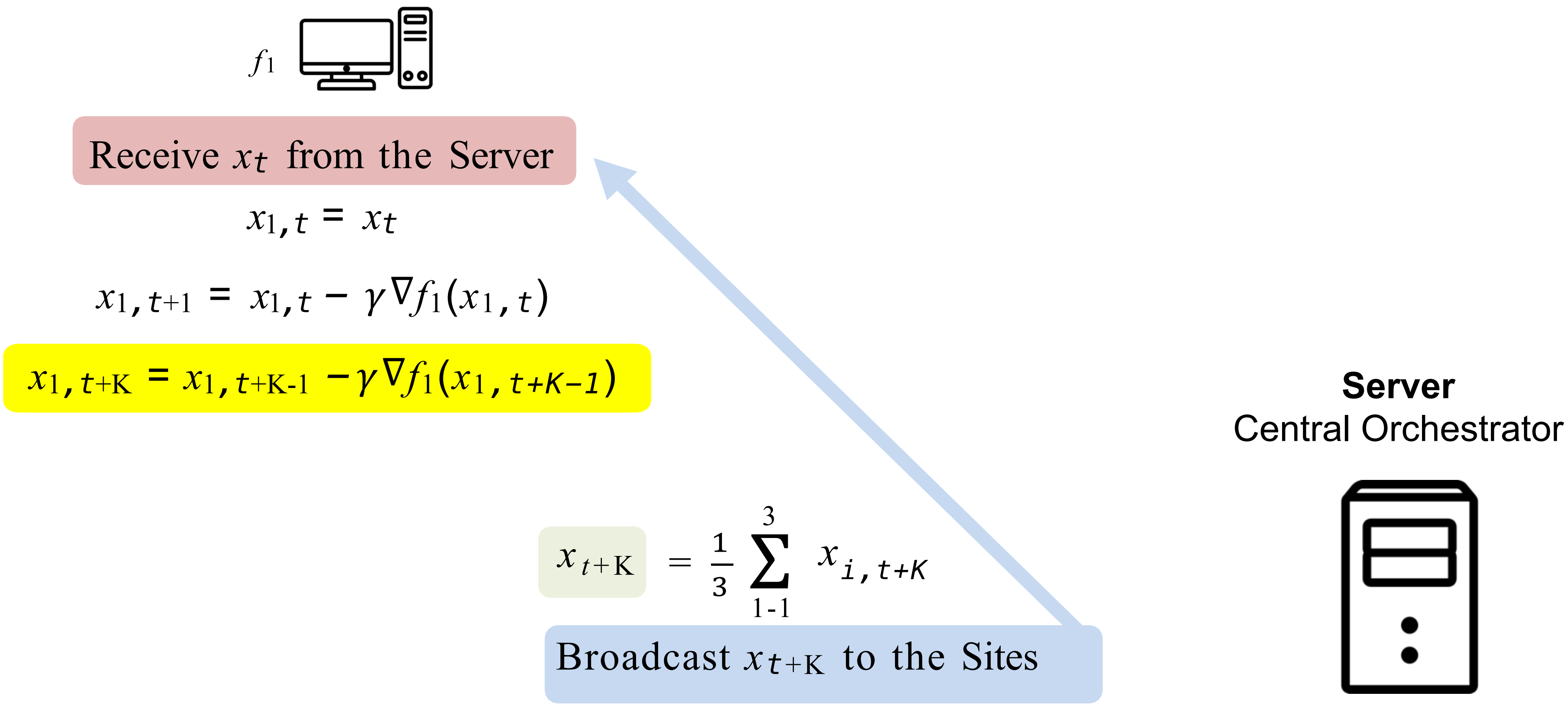
Server

$$x_{t+1} = \frac{1}{3} \sum_{1=1}^{3} x_{i,t+1} =$$

Broadcast x_{t+1} to the Sites

Distributed Local Gradient Descent

Site 1,2,3 – Computer, Mobile Device, Hospital



f_1

Receive x_t from the Server

$$x_{1,t} = x_t$$

$$x_{1,t+1} = x_{1,t} - \gamma \nabla f_1(x_{1,t})$$

$$x_{1,t+K} = x_{1,t+K-1} - \gamma \nabla f_1(x_{1,t+K-1})$$

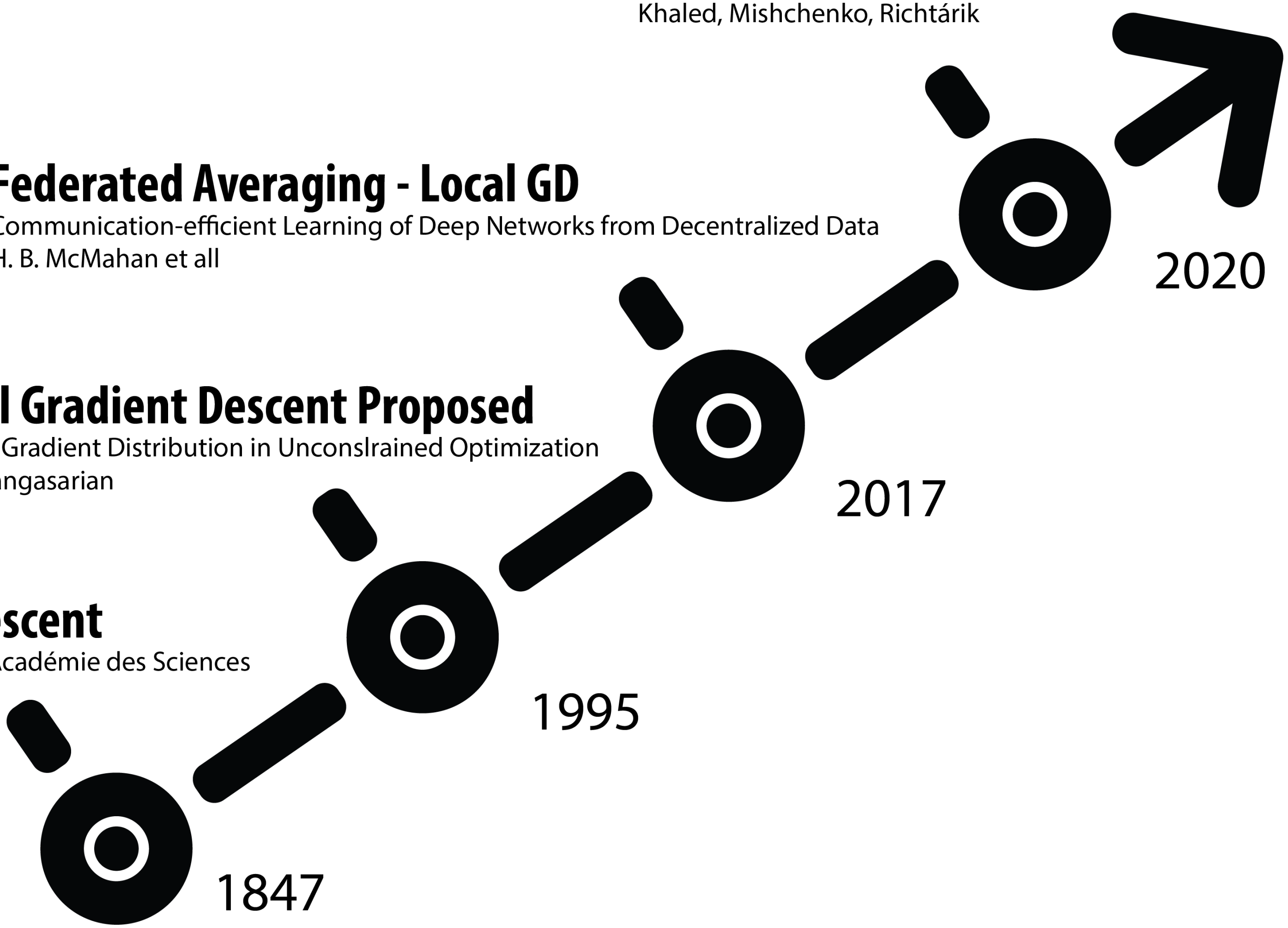
$$x_{t+K} = \frac{1}{3}\sum_{1-1}^{3} x_{i,t+K}$$

Broadcast x_{t+K} to the Sites

Server
Central Orchestrator

Part 3
Brief History of Local Training

From Gradient Descent to Local Gradient Descent



Gradient Descent

Compte Rendu á l'Académie des Sciences
L. A. Cauchy

1847

Local Gradient Descent Proposed

Parallel Gradient Distribution in Unconslrained Optimization
O. L. Mangasarian

1995

Federated Averaging - Local GD

Communication-efficient Learning of Deep Networks from Decentralized Data
H. B. McMahan et all

2017

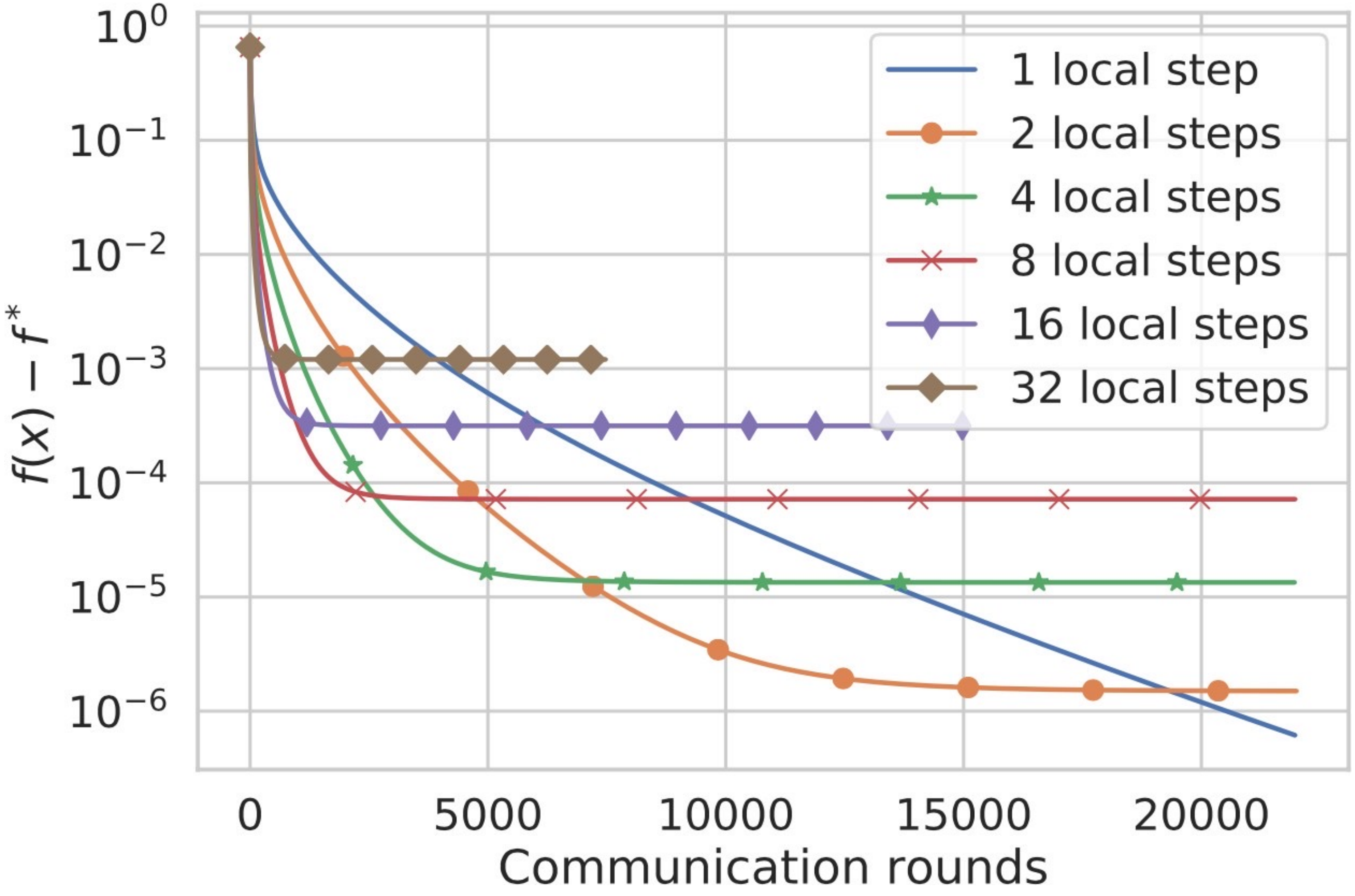
First general Theory for Local GD

First Analysis of Local GD on Heterogeneous Data
Khaled, Mishchenko, Richtárik

2020

Part 4
What does Local Training do?

Local Training



Thank You